Developing NASA-TLX – 1986-2010

Lowell Staveland, SHFE Inc.

Who I Am

I. From Davis to NASA, Anthropology to Psychology.

II. From TLX to TLM

III. From NASA to Industry

IV. Back to NASA

Scale development

I. Process - Years of Regression Analyses

- Some people spend years in therapy to figure out what’s happening in their own heads. We spent years in Regression figuring out what was happening in other’s heads.
- Started with 19 workload factors, reduced to 14 then to 10 (9 component + 1 Overall)
- Conducted 25 tests with 10 scales, used results from 16 to derive the TLX 6 scales and 6 weights.
- The tests were grouped into 6 categories of experimental conditions with different primary sources of loading.

✓	Cognitive Load -	Simple Discrete Tasks
✓	Manual load –	Single axis manual control
✓	Cognitive and Manual load –	Dual tasks
✓	Response Selection vs Execution loads -	FittsBerg Tasks
✓	Temporal Loads –	Supervisory Control Tasks
✓	Load Complexity and Difficulty –	Simulated Flight Tasks

Scale development

II. What do I remember –

- Conducted many tests followed by many regressions. (including many Non-parametric Komalgorov-Schmirnoff tests – not used this one since)
- Lots of meetings to map out test WL assessments, discussing correlations, circling numbers, erasing, re-circling.
- Used the latest results to predict WL and performance correlations with different factors. Followed this with all possible options not covered by predictions
- Retested predictions during next tests. Repeated cycle with set of 10 scales.
    - Dropped scales after 1 or more tests in which they really did not contribute to the variance in OW or in-consistently contributed. Consistence defined by beta weights and correlations of < %50.
    - Wrote programs to run regressions and learned to use Apple IIE, vi and Unix printing commands.

III. Sources of Workload

- We assumed WL was multi-dimensional with different sources - either task based or personal.
- Constructed tests to manipulate the loading source and determine WL.

Scale Development

IV. Scale Definitions

- Altered wording of the definitions to be concept specific and task independent
- Subsequent research suggests changing Scale Definitions to include test task and concept specificity improve scales capability to capture “actual workload”.

V. Scales and Weights: Current *[Previous]*

- Performance: Mental, Physical, & Temporal Demand *[Task Difficulty, Time Pressure and Activity Type]*
- Behavioral: Effort, Performance *[Physical, Mental, Own Performance]*
- Subjective: Frustration *[Frustration, Stress, Fatigue]*
- Composite: Weighted WL, [*Overall Workload]*

Scale Development

VI. Scale Structure

- Tried different intervals from 100 pt to 10 pt and found 0-10 pt .5 interval (20pt) was as effective as more intervals.
- Using computer, paper/pencil or verbal didn’t seem to have a big affect.
- A line with anchors, and interval marks without numbers seemed best.

VII. Weights

- Weights reflect variation in the sources of tasks load, and reduce between-subject variability of ratings, when taken AFTER the task.
- We initially put a lot of “weight” on weighting ratings - didn’t think the WL ratings could stand on their own, therefore less accepted.
- As it turns out, we weren’t quite right using weights.
  - ✓ Ratings hold up on their own without be weighted.
  - ✓ Weights became another separate diagnostic measure of the source of demands even if not as sensitive to varying demands.

24 Years of Use

VIII.Ms. Hart conducted a survey in 2006 of 550 studies in which TLX was used or reviewed and found it's been –

a. reasonably easy to use and reliably sensitive to experimentally important manipulations.

b. translated into more than a dozen languages, administered verbally, in writing, or by computer, and modified in a variety of ways.

c. subjected to a number of independent evaluations in which its reliability, sensitivity, and utility were assessed and compared to other methods

d. Used on all continents except Antarctica, primarily in N America and Europe by Government Organizations and Universities.

e. In a wide range of operational environments targeting interface design or evaluation, systems control, teamwork, SA, flying, driving, monitoring, communications.

f. Still subject to same methodological issues with context and anchor effects, inter-correlations and redlines.

g. Change to meet needs – modify 3 of scales, no weights (Raw-TLX), use component scales individually- weighted and unweighted.

Future Uses/Development

I. Combine of Scales

   a. Currently different combinations of scales depending on needs: Workload with Handling qualities, numerical with comparative, indirect with direct.

II. Collect Ratings Real-Time

   a. Call out only relevant ratings to simplify.

   b. Complete others retrospectively

III. Rate the timeline

   a. Instead marking a scale for a task, create corresponding rating “Scale-lines’” with estimates for relevant events.

IV. Combine Rating with Verbal Protocols

   a. Debrief each rating point to get reasons: display at t3 not useful.

V. Use Ratings to find drill down points

   a. Use ratings to discover points of interest to investigate with additional tools or finer grain level.